**Programming (bagging).**

**Overview**

This program is an extension of Homework 1 in which ID3 Decision tree algorithm is implemented. This program takes the data files to build different decision trees. Hence we have N number of Training Models if N training files are given as input

Program takes one test file as a 2nd argument whose class label is evaluated based on the decision made by N different training Models.

**Bagging:**

Here, the first argument of the program is the directory containing N different datasets. All the datasets must consistently have same number of attributes i.e Random Forests are not allowed meaning if One training set has 4 attributes, all the other training Data sets must also have 4 attributes.

N different decision trees are constructed based on the N different data sets. Then the trained trees are used to determine the class label for each record of the test data. Based on the majority votes by N different trees, class label is assigned for the test data.

Finally the assigned label is compared to actual label to determine the accuracy of Bagging.

**Algorithm/Pseudo Code**

Step 1: For each training data construct the ID3 decision tree

Step 2: Consider each record of the test data and traverse the tree based on the attribute values of the record until the leaf node is reached.

Step 3: Evaluate the class label based on the leaf node’s classification label.

Step 4. Repeat step 2 and 3 for all the ID3 Decision trees constructed.

Step 5: Based on the majority votes of the trees, classify the output label of the record.

Step 6. Repeat step2 to step 5 for all the records of the test data.

Step 7. Compare the label assigned by the Bagging model with the actual label to determine the accuracy.

**Observation and Experimental result:**

It has been observed that Bagging would help to certain extent in improving the accuracy. However the results observed will become constant after reaching a threshold value.

The number and type of records that are chosen to create the training files would affect the accuracy i.e If the number of records are chosen from all the section of the training file improves accuracy compared to the records chosen from one part of the training file.

The following table provides the experimental results for the different number of bagging

|  |  |
| --- | --- |
| Number of Bagging | Accuracy for the Test Data(in %) |
|  |  |
| 1 | 77.339 |
| 10 | 79.339 |
| 25 | 86.69 |
| 50 | 86.69 |
| 75 | 86.69 |

As we can see in the table when only one train data is given we have an accuracy of about 77.4% whrere as accuracy increases to 86.69% after bagging it with 25 training models and remains same thereafter.

**Conclusion:**

Bagging is always preferable over classification using single classifier. However, bagging does not improve once the stable state is reached.